Artificial Theory of Mind and our Collective Intelligence:

Exploring the social gap between humans and machines

GSO 2025, Tübingen, Germany

Michael Harré, Jaime Ruiz-Serra, Catherine Drysdale

Modelling and Simulation Group

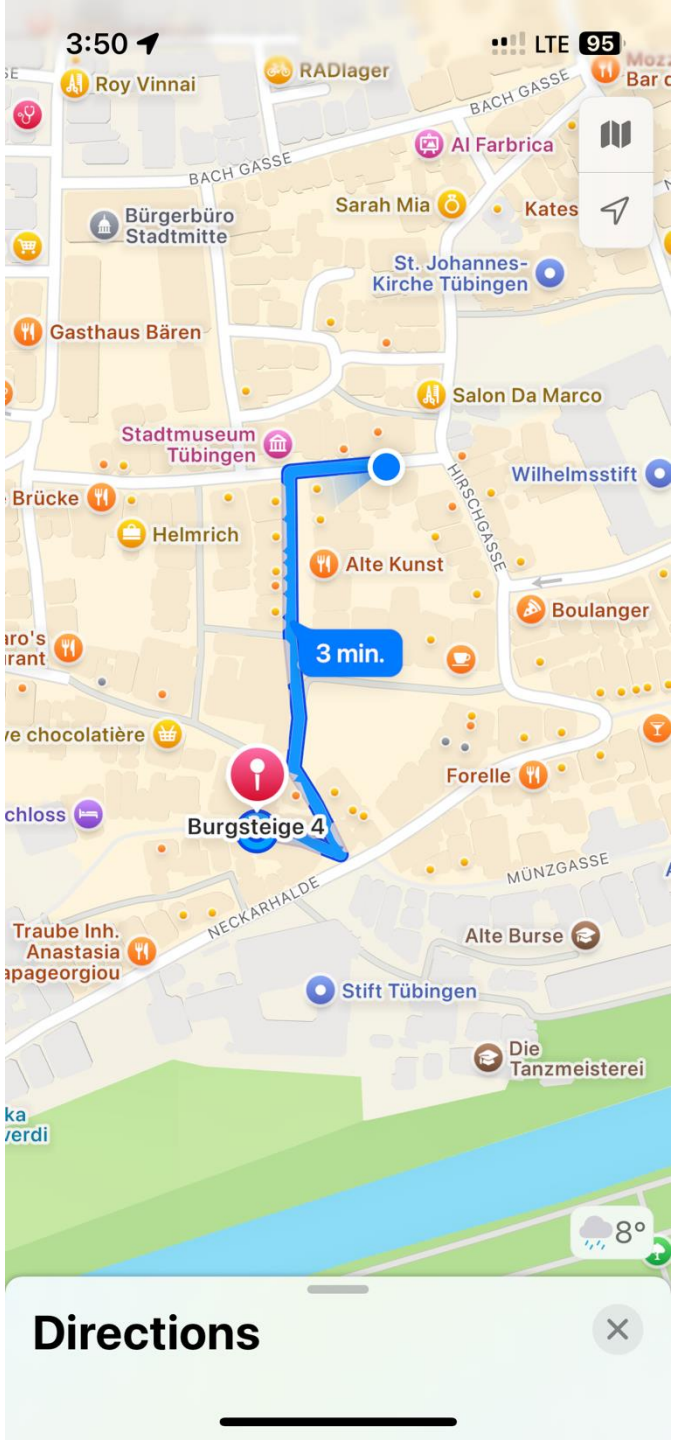
School of Computer Science

University of Sydney

THE UNIVERSITY OF SYDNEY

https://enterprisetalk.com/tech-trends/ai/human-intelligence-can-fix-ai-shortcomings/

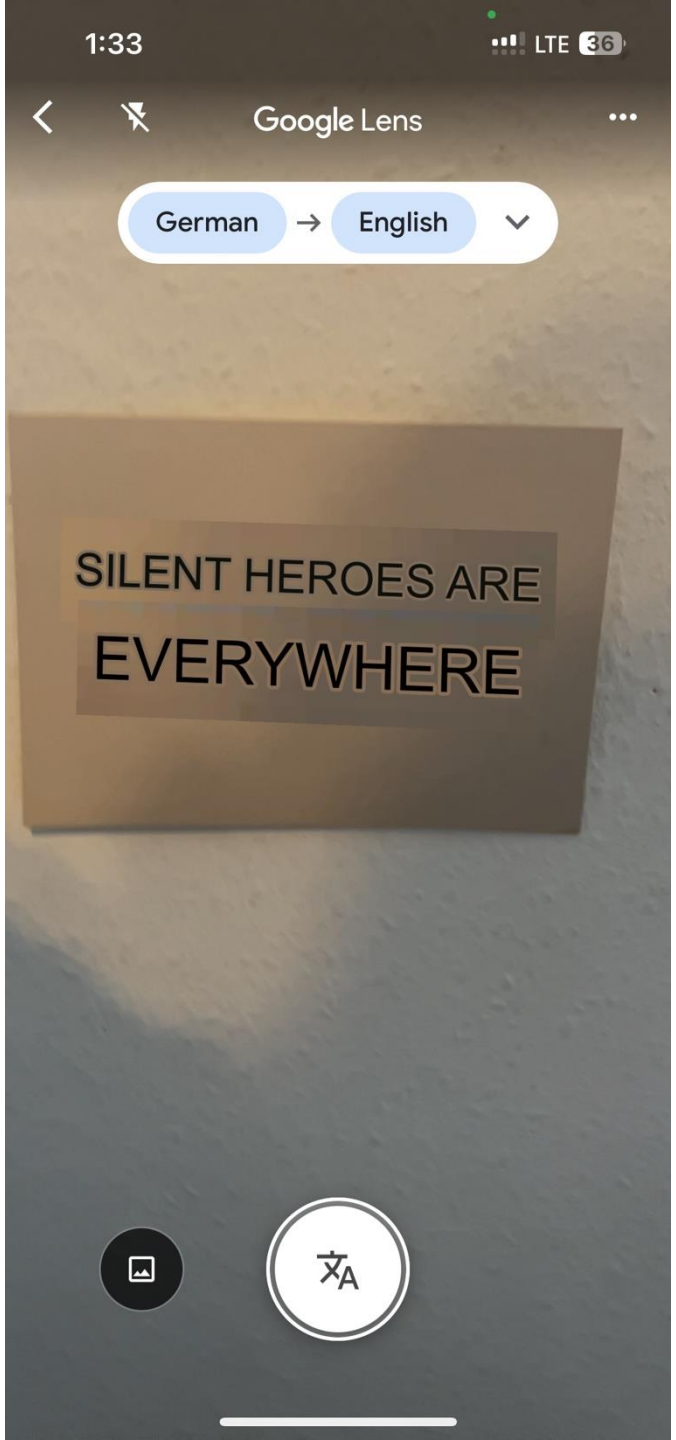
Overview


3:50
LTE 95
Roy Vinnai
RADlager
BACH GASSE
Al Farbrica
Sarah Mia
Kates
Bürgerbüro Stadtmitte
St. Johannes-Kirche Tübingen
Gasthaus Bären
Salon Da Marco
Stadtmuseum Tübingen
Wilhelmsstift
Brücke
Helmrich
HIRSCHGASSE
Alte Kunst
Boulanger
3 min.
Forelle
Burgsteige 4
MÜNZGASSE
NECKARHALDE
Traube Inh. Anastasia
Alte Burse
Stift Tübingen
Die Tanzmeisterei
8°
Directions

Overview


STILLE HELDEN
SIND ÜBERALL
0.5
1×
CINEMATIC
VIDEO
PHOTO
PORTRAIT
PANO

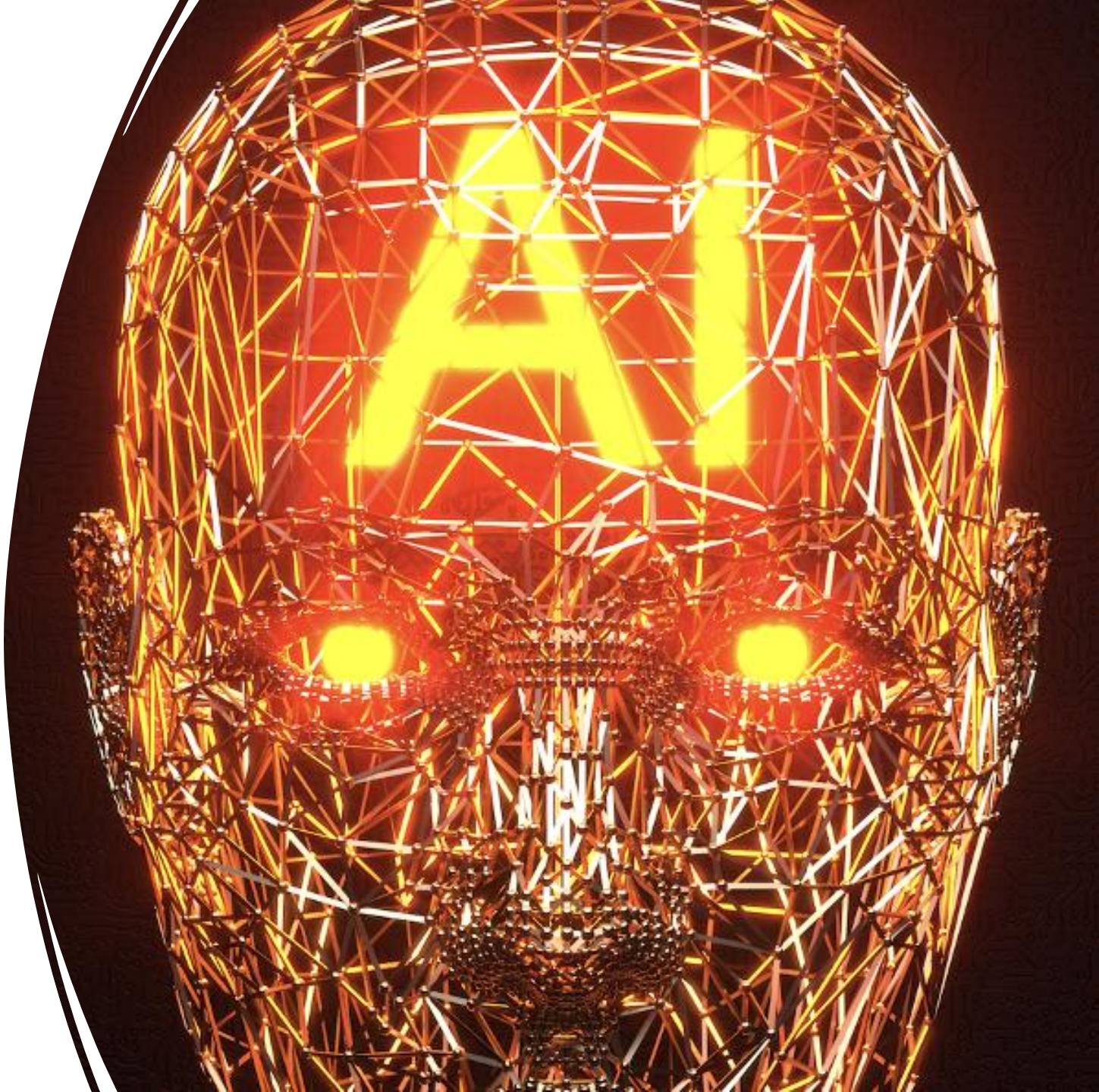
Overview


1:33
LTE 36
Google Lens
German
English
SILENT HEROES ARE
EVERYWHERE

Overview

"I think there is a world market for maybe five computers."

Thomas Watson, president of IBM, 1943



Overview

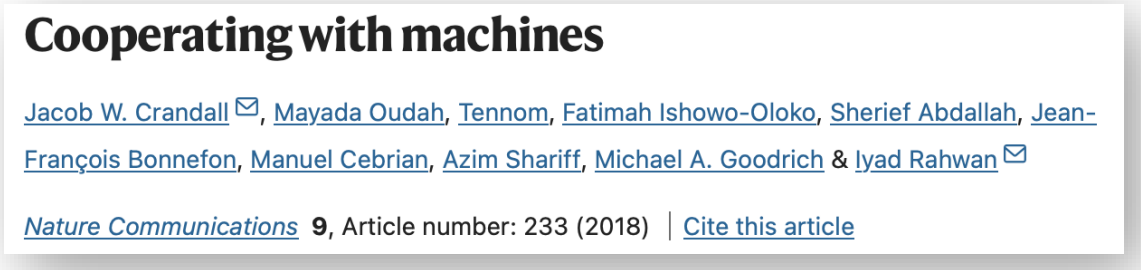
"I think there is a world market for maybe five computers."

Thomas Watson, president of IBM, 1943

Overview

Overview

- **Key Questions:**
  - How can multiple AI agents coordinate to achieve goals?
  - How can AI interact with humans effectively?
  - What tools can we use?
- **Grand Challenge (Garibay et al.):**
  - Respecting human cognitive processes at the human–AI interaction frontier
- **Core Claim:**
  - This also extends to AI–AI interaction;
  - Based on human psychology as the only working model we have



Overview

- **Key Questions:**
  - How can multiple AI agents coordinate to achieve goals?
  - How can AI interact with humans effectively?
  - What tools can we use?

- **Grand Challenge (Garibay et al.):**
  - Respecting human cognitive processes at the human–AI interaction frontier

- **Core Claim:**
  - This also extends to AI–AI interaction;
  - Based on human psychology as the only working model we have

Overview

- **Key Questions:**
  - How can multiple AI agents coordinate to achieve goals?
  - How can AI interact with humans effectively?
  - What tools can we use?

- **Grand Challenge (Garibay et al.):**
  - Respecting human cognitive processes at the human–AI interaction frontier

- **Core Claim:**
  - This also extends to AI–AI interaction;
  - Based on human psychology as the only working model we have

Overview

- **Key Questions:**
  - How can multiple AI agents coordinate to achieve goals?
  - How can AI interact with humans effectively?
  - What tools can we use?

- **Grand Challenge (Garibay et al.):**
  - Respecting human cognitive processes at the human–AI interaction frontier

- **Core Claim:**
  - This also extends to AI–AI interaction;
  - Based on human psychology as the only working model we have

Overview

- **Key Questions:**
  - How can multiple AI agents coordinate to achieve goals?
  - How can AI interact with humans effectively?
  - What tools can we use?

- **Grand Challenge (Garibay et al.):**
  - Respecting human cognitive processes at the human–AI interaction frontier

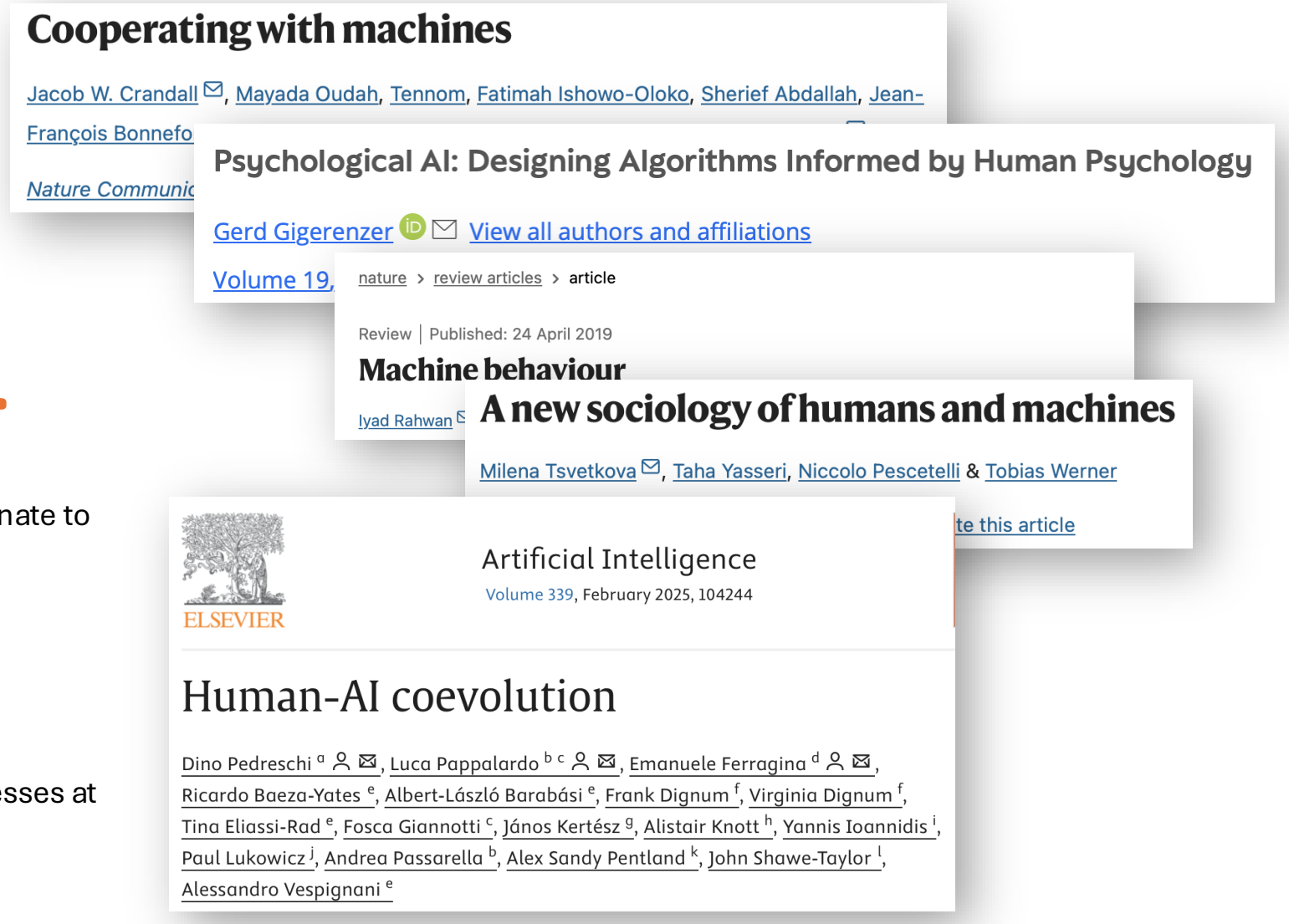
- **Core Claim:**
  - This also extends to AI–AI interaction;
  - Based on human psychology as the only working model we have

Overview

- **Key Questions:**
  - How can multiple AI agents coordinate to achieve goals?
  - How can AI interact with humans effectively?
  - What tools can we use?

- **Grand Challenge (Garibay et al.):**
  - Respecting human cognitive processes at the human–AI interaction frontier

- **Core Claim:**
  - This also extends to AI–AI interaction;
  - Based on human psychology as the only working model we have

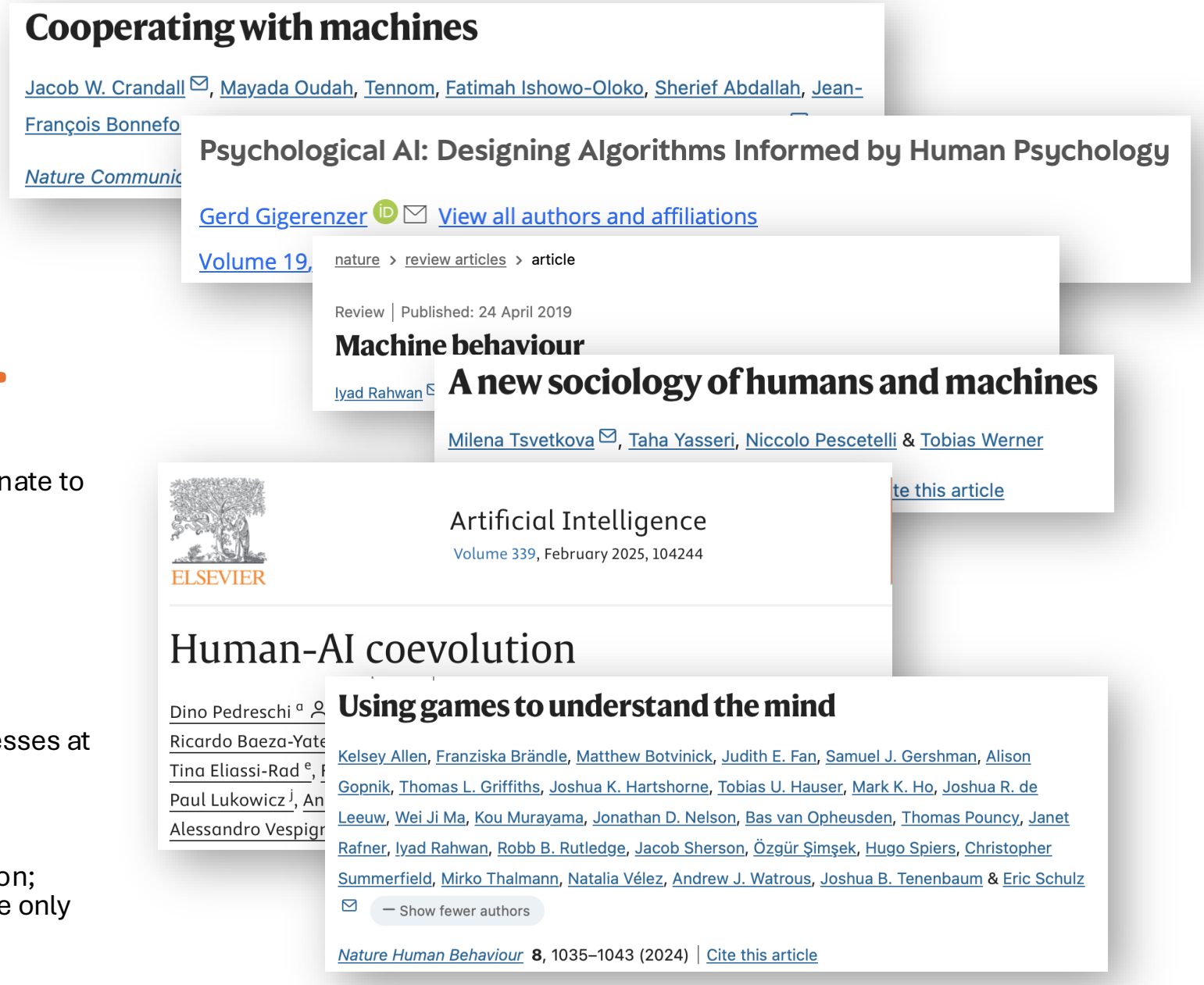


Overview

- **Key Questions:**
  - How can multiple AI agents coordinate to achieve goals?
  - How can AI interact with humans effectively?
  - What tools can we use?

- **Grand Challenge (Garibay et al.):**
  - Respecting human cognitive processes at the human–AI interaction frontier

- **Core Claim:**
  - This also extends to AI–AI interaction;
  - Based on human psychology as the only working model we have

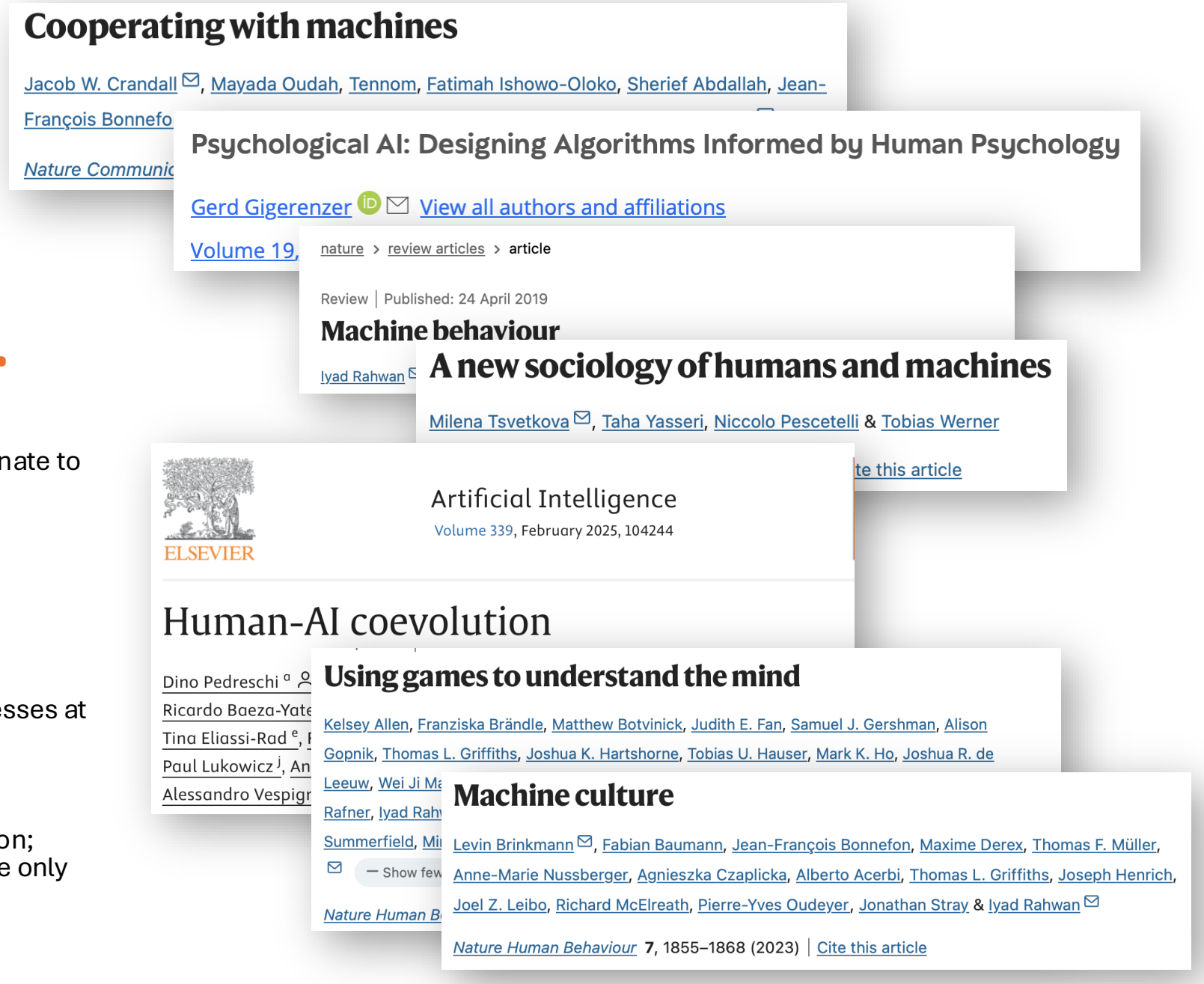


Overview

- **Key Questions:**
  - How can multiple AI agents coordinate to achieve goals?
  - How can AI interact with humans effectively?
  - What tools can we use?

- **Grand Challenge (Garibay et al.):**
  - Respecting human cognitive processes at the human–AI interaction frontier

- **Core Claim:**
  - This also extends to AI–AI interaction;
  - Based on human psychology as the only working model we have



Approach

Biological Inspirations:

- Single neuron complexity in neural networks
- Ant colony adaptability and social organisation

https://www.brainfacts.org/brain-anatomy-and-function/cells-and-circuits/2022/receptors-and-molecular-signaling-113022

Approach

Biological Inspirations:

- Single neuron complexity in neural networks
- Ant colony adaptability and social organisation

https://stockcake.com/i/ants-in-interaction_922626_532926

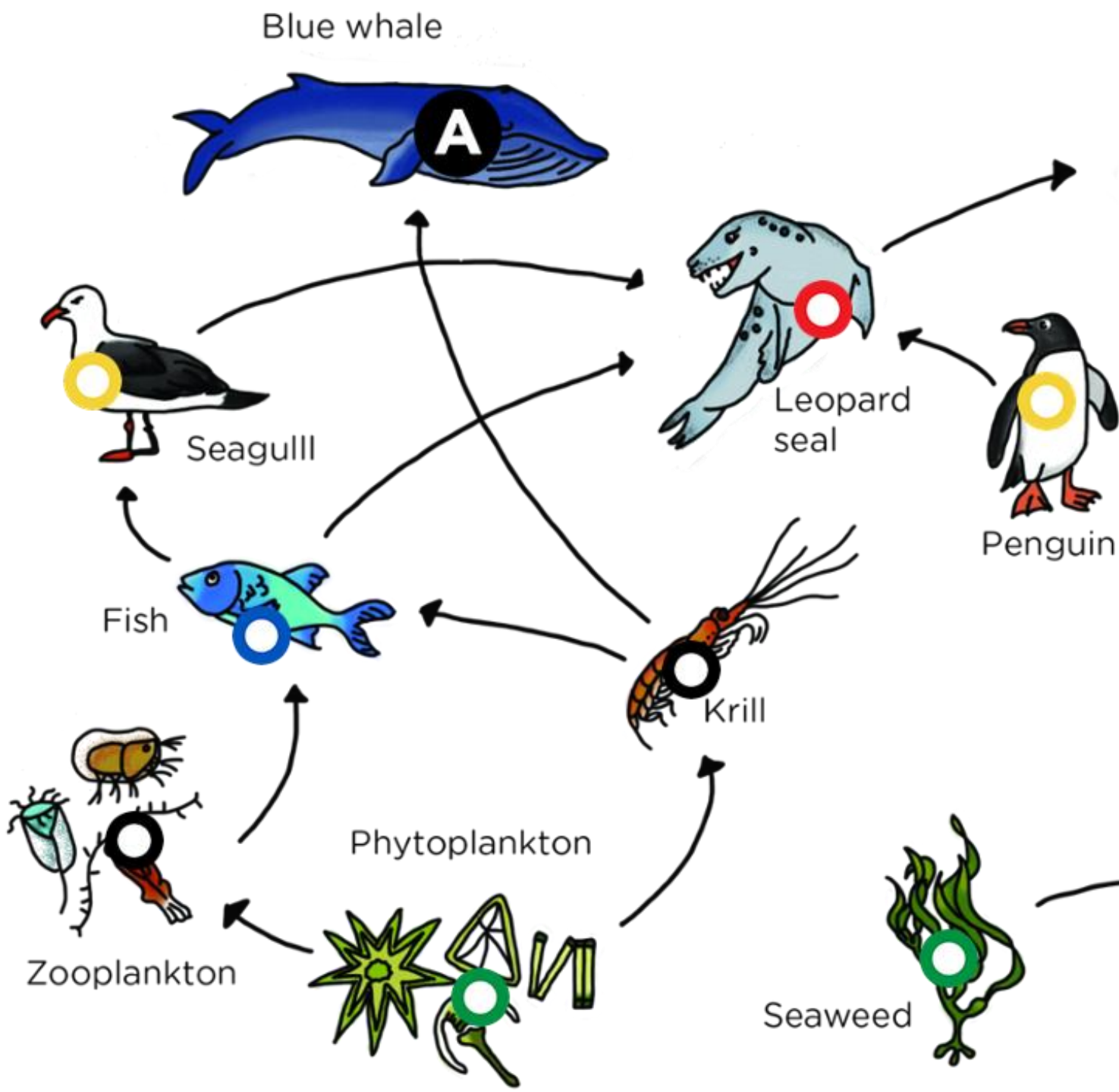
Approach

Biological Inspirations:

- Single neuron complexity in neural networks
- Ant colony adaptability and social organisation

Ecological Analogy:

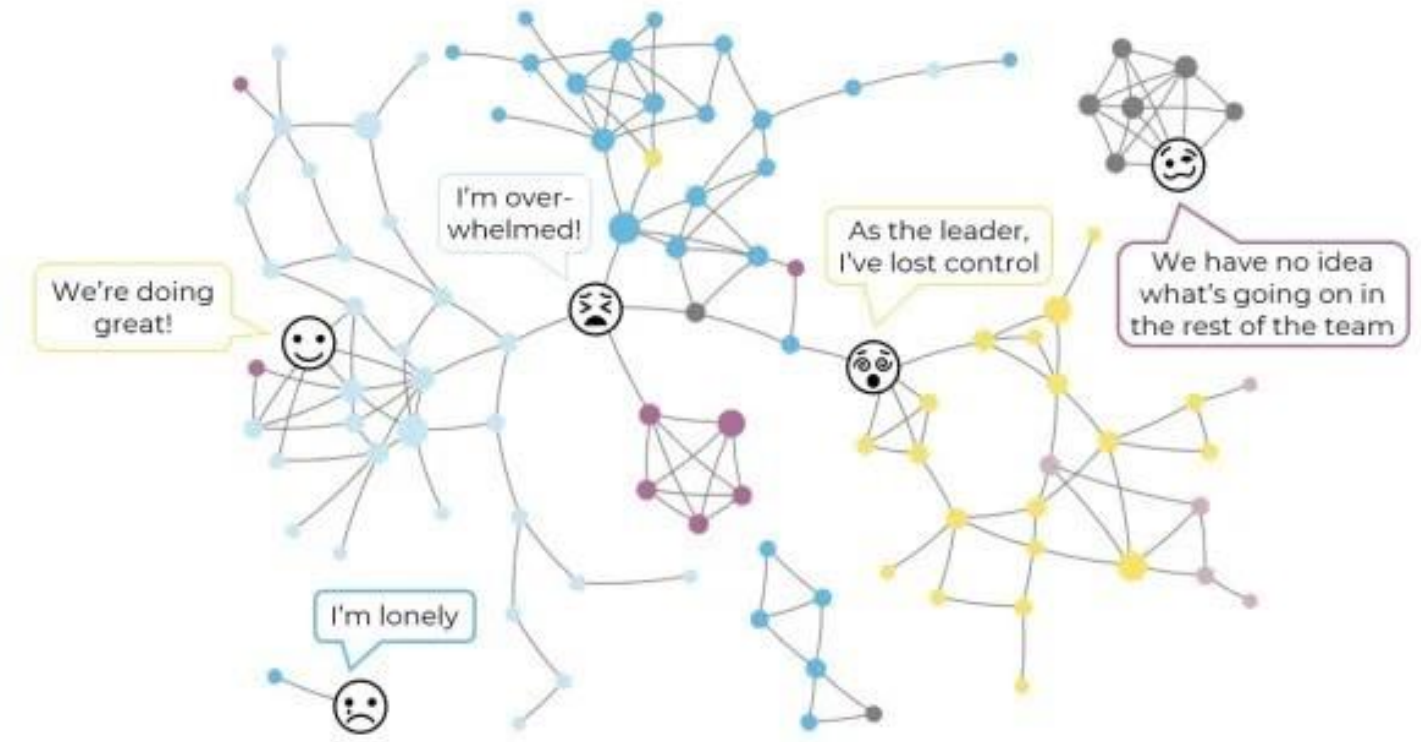
- Niche choice,
- Niche construction,
- And niche conformity



Approach

Human Social Networks:

- Influenced by neurophysiology, psychology, language
- Theory of Mind (ToM) as a central coordinating faculty



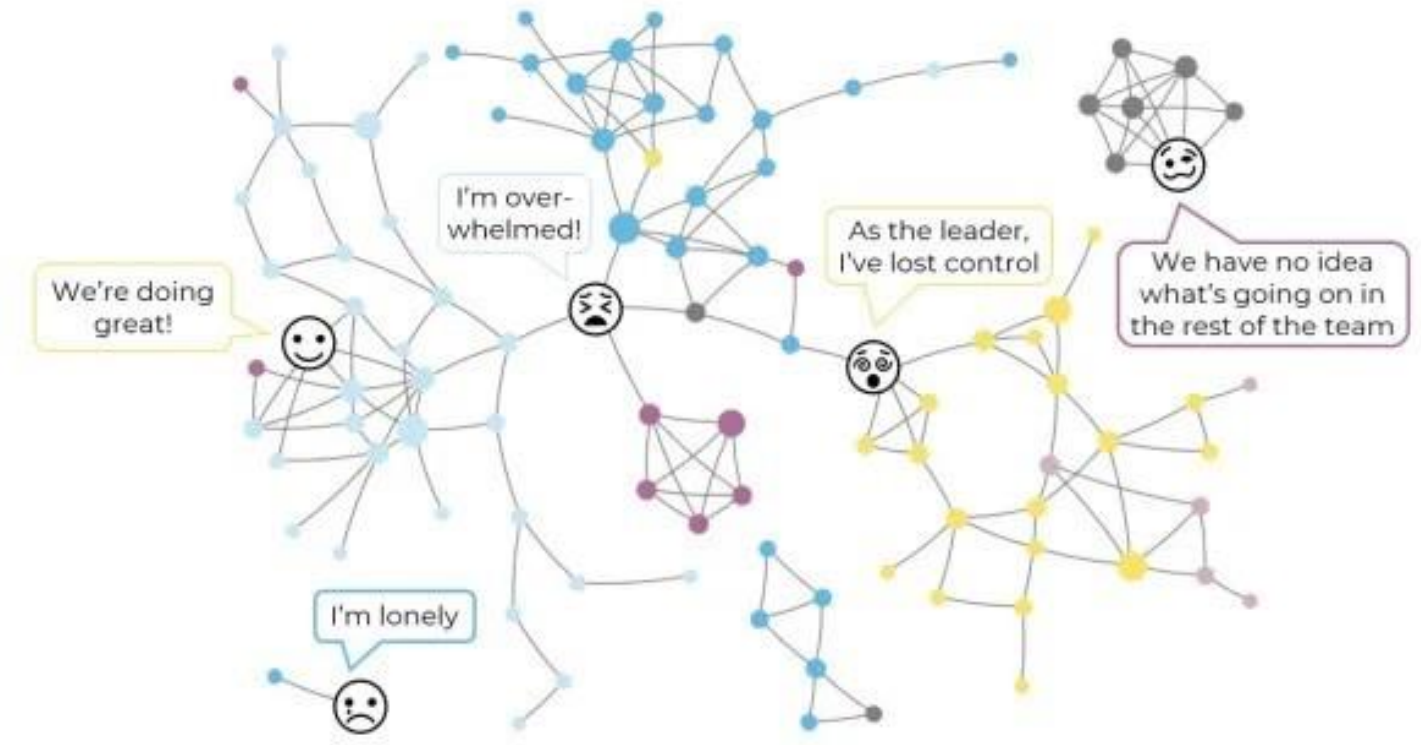
Approach

Human Social Networks:

- Influenced by neurophysiology, psychology, language
- Theory of Mind (ToM) as a central coordinating faculty

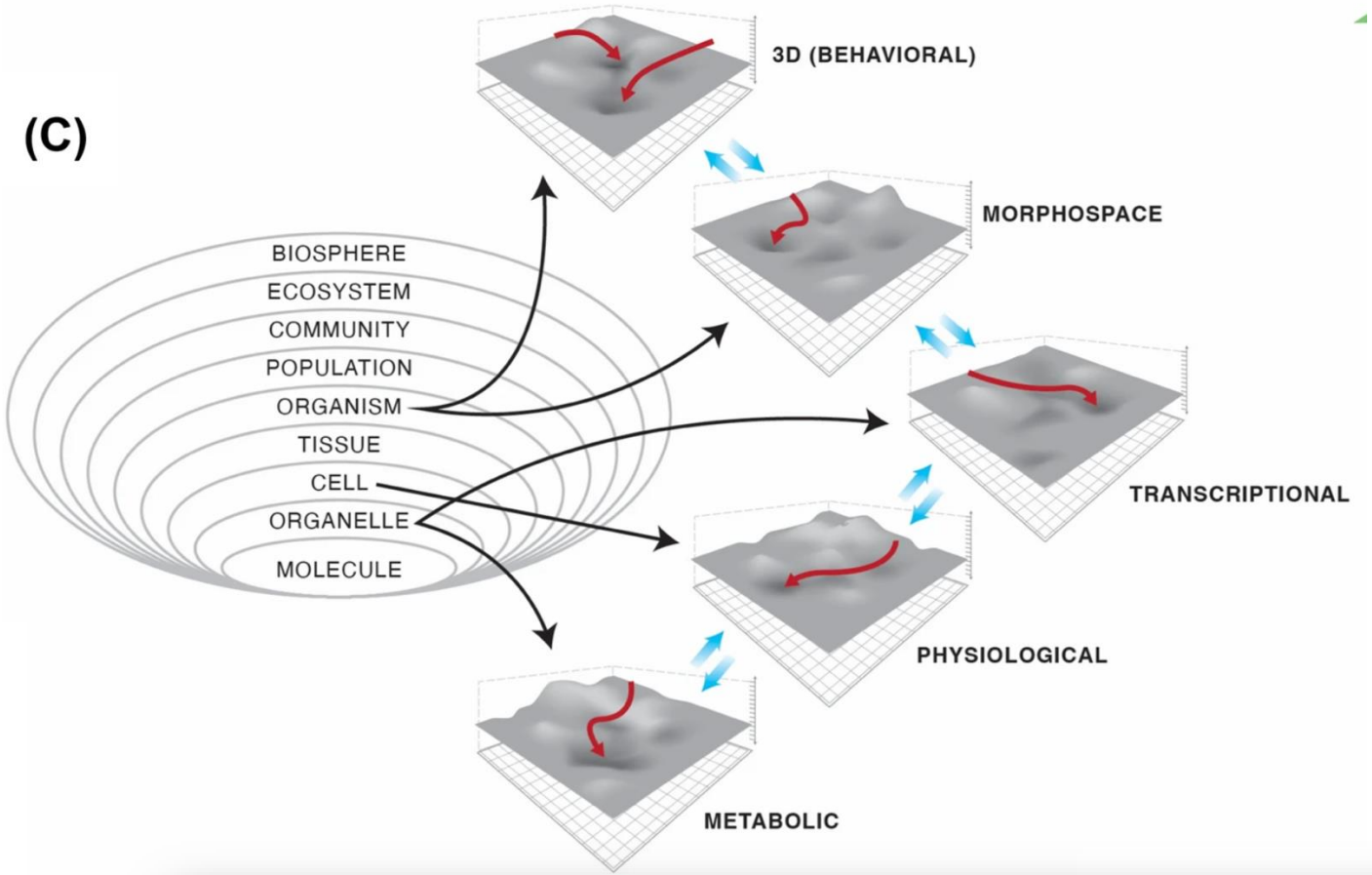
Goal:

- Outline a path toward a collective AI that can participate in and help guide social structures



Single Neuron Complexity

Biological Agents: Composed of organs, cells, molecular networks



Bioelectric networks: the cognitive glue enabling evolutionary scaling from physiology to mind

Michael Levin

Review | Open access | Published: 19 May 2023

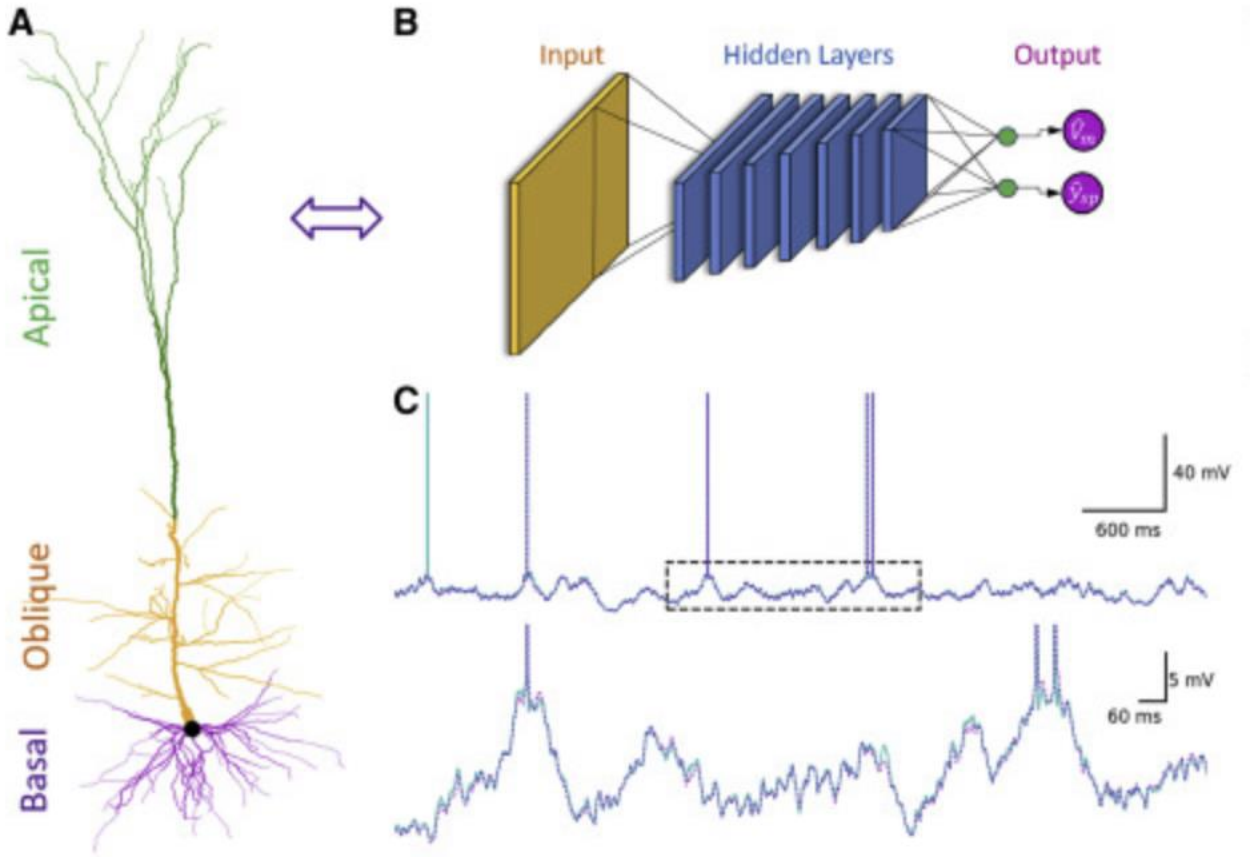
Volume 26, pages 1865–1891, (2023) Cite this article

Single Neuron Complexity

Biological Agents: Composed of organs, cells, molecular networks

Neuron "competencies" (Levin):

- Beniaguev et al.: One neuron ≈ 5–8 layers of a deep network
- Complex adaptation: Each neuron as an "agent"



ARTICLE · Volume 109, Issue 17, P2727-2739.E3, September 01, 2021 · *Open Archive*

Download Full Issue

Single cortical neurons as deep artificial neural networks

David Beniaguev [1,3] · Idan Segev [1,2] · Michael London [1,2]

Affiliations & Notes · Article Info

Single Neuron Complexity

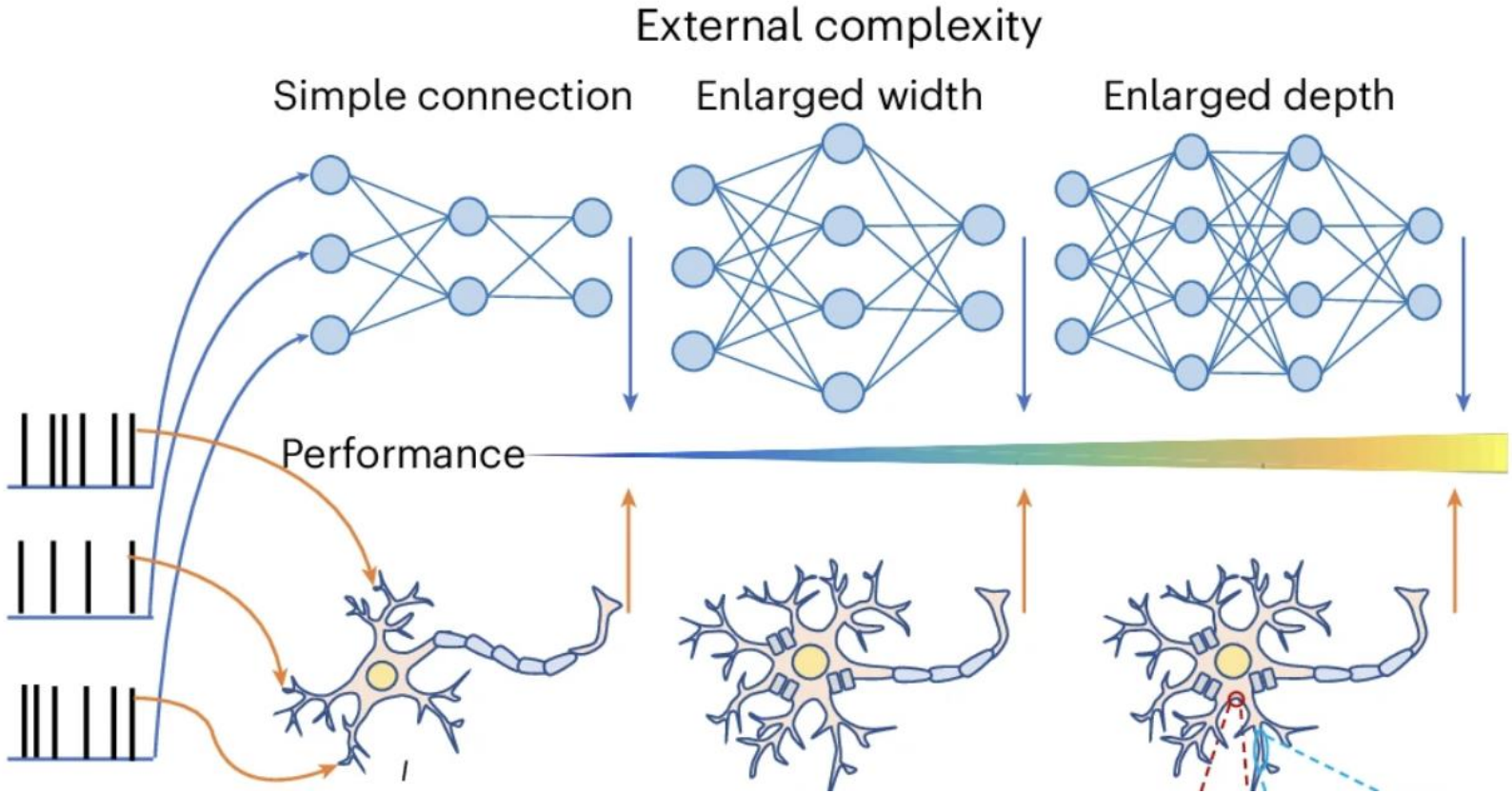
Biological Agents: Composed of organs, cells, molecular networks

Neuron "competencies" (Levin):

- Beniaguev et al.: One neuron ≈ 5–8 layers of a deep network
- Complex adaptation: Each neuron as an "agent"

Implication for AI:

- These neural agents are highly sophisticated at small scales
- AI ANNs + Human Bio-NNs can be combined to produce a "collective intelligence"



Article | Published: 16 August 2024

Network model with internal complexity bridges artificial intelligence and neuroscience

Linxuan He, Yunhui Xu, Weihua He, Yihan Lin, Yang Tian, Yujie Wu, Wenhui Wang, Ziyang Zhang, Junwei Han, Yonghong Tian, Bo Xu & Guoqi Li

Nature Computational Science **4**, 584–599 (2024) | Cite this article

Ant Colony Adaptability

Conserved Social Structures:

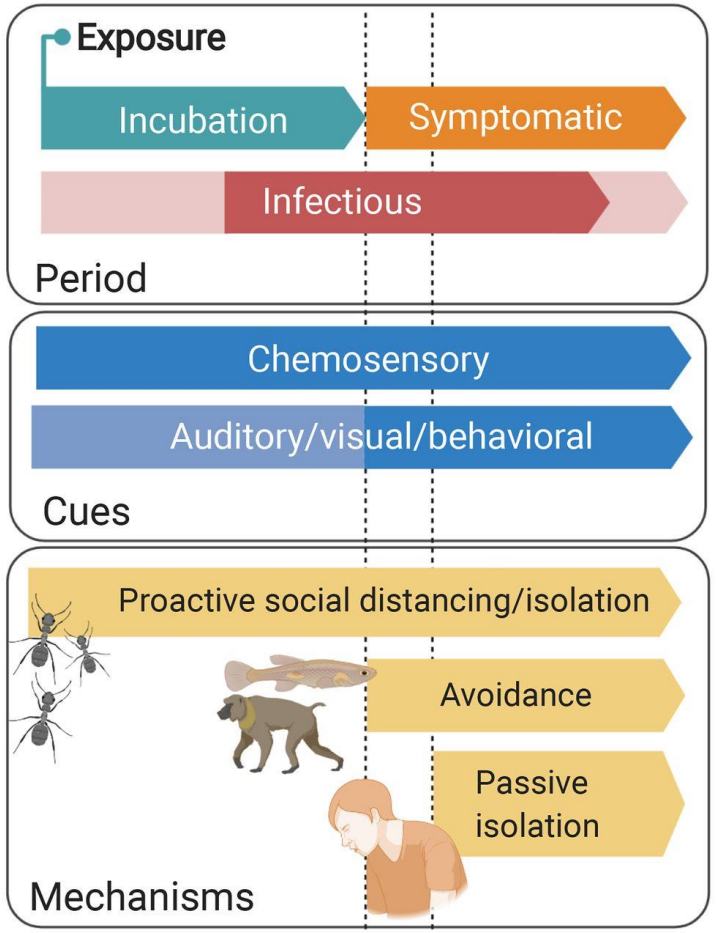
- Persist over 100+ million years of evolution

Division of Labor & Leadership:

- Improves collective performance

Adaptation to threats:

- Reorganize social behaviours
- Modify nests to curb disease



Infectious diseases and social distancing in nature

SEBASTIAN STOCKMAIER, NATHALIE STROEYMEYT, ERIC C. SHATTUCK, DANA M. HAWLEY, LAUREN ANCEL MEYERS, AND DANIEL I. BOLNICK

Authors Info & Affiliations

SCIENCE • 5 Mar 2021 • Vol 371, Issue 6533 • DOI: 10.1126/science.abc8881

Ant Colony Adaptability

Conserved Social Structures:

- Persist over 100+ million years of evolution

Division of Labor & Leadership:

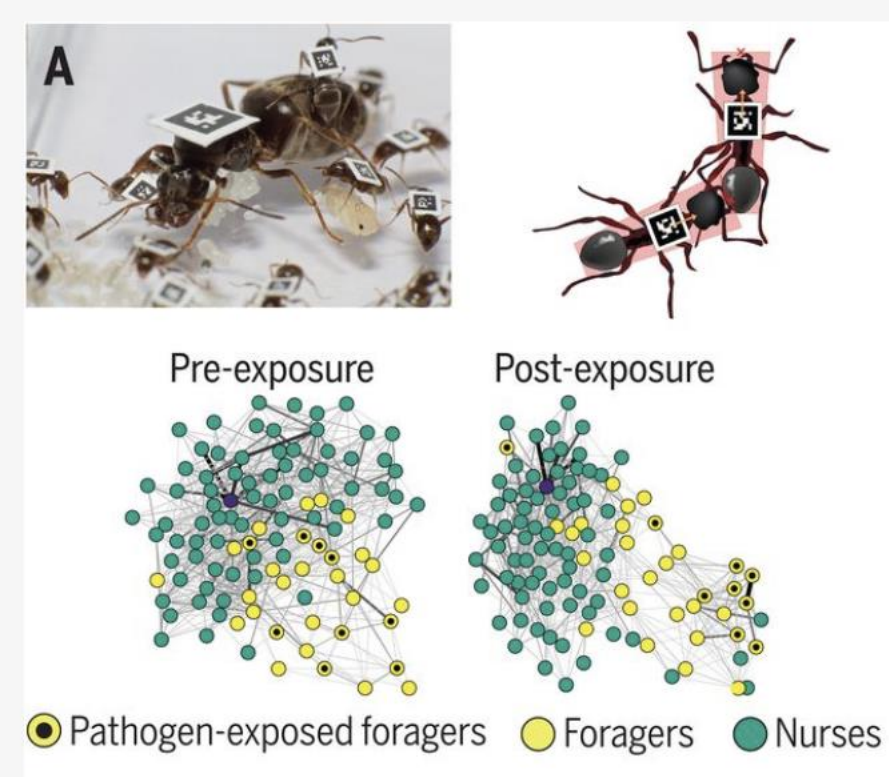
- Improves collective performance

Adaptation to threats:

- Reorganize social behaviours
- Modify nests to curb disease

Insight for AI:

- Fluid, dynamic roles enhance group dynamics



Infectious diseases and social distancing in nature

SEBASTIAN STOCKMAIER, NATHALIE STROEYMEYT, ERIC C. SHATTUCK, DANA M. HAWLEY, LAUREN ANCEL MEYERS, AND DANIEL I. BOLNICK

Authors Info & Affiliations

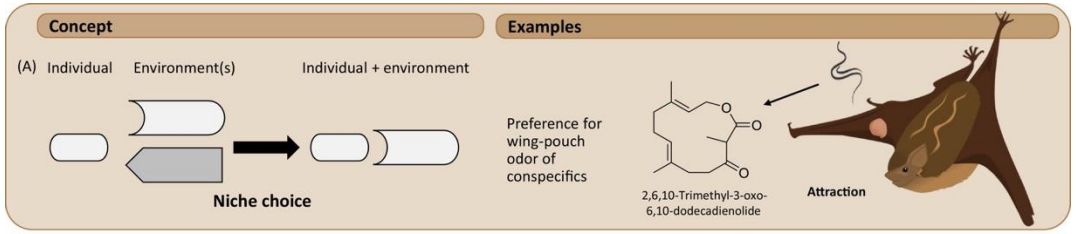
SCIENCE • 5 Mar 2021 • Vol 371, Issue 6533 • DOI: 10.1126/science.abc8881

Ecological Networks

From Ecology to Social Organisation:

- Individuals interact with an environment to form “niches”

Ecological Networks



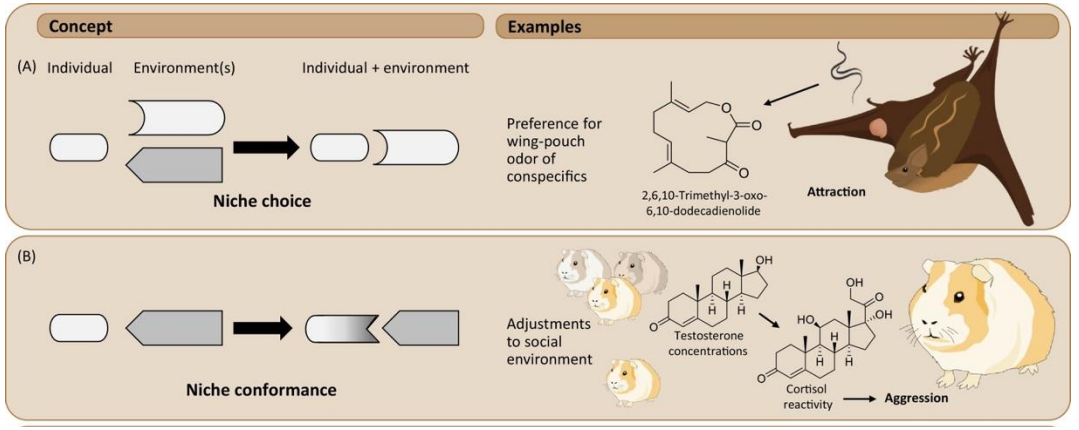
From Ecology to Social Organisation:

- Individuals interact with an environment to form “niches”

Main Processes (Odling-Smee, Laland, etc.):

- Niche Choice: Selecting fitting environmental conditions

Ecological Networks



From Ecology to Social Organisation:

- Individuals interact with an environment to form “niches”

Main Processes (Odling-Smee, Laland, etc.):

- Niche Choice: Selecting fitting environmental conditions
- Niche Conformance: Adjusting phenotype to the environment

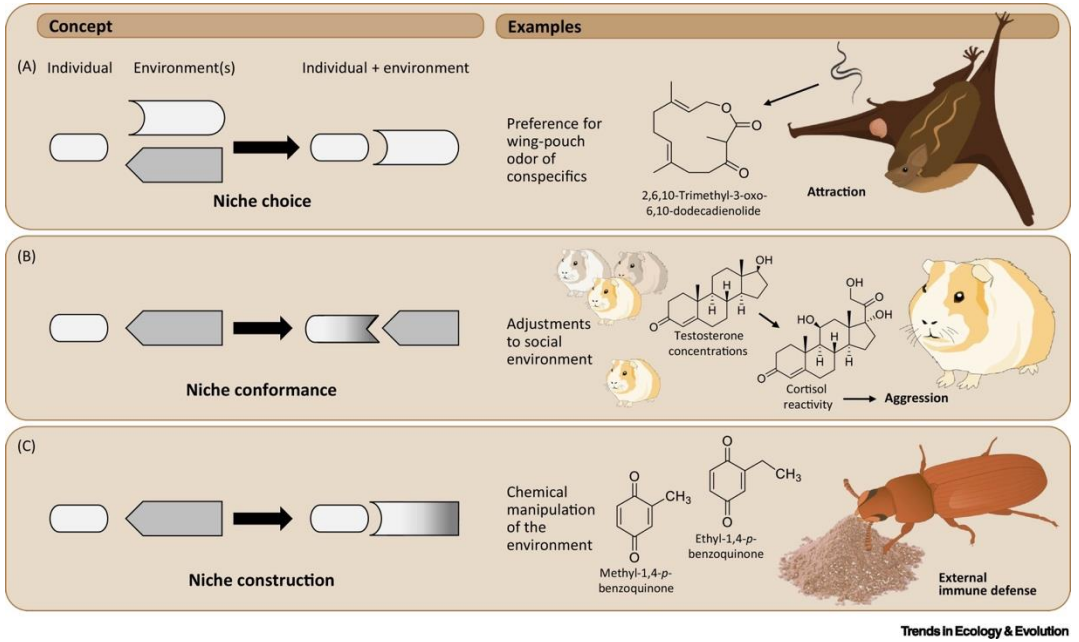
Ecological Networks

From Ecology to Social Organisation:

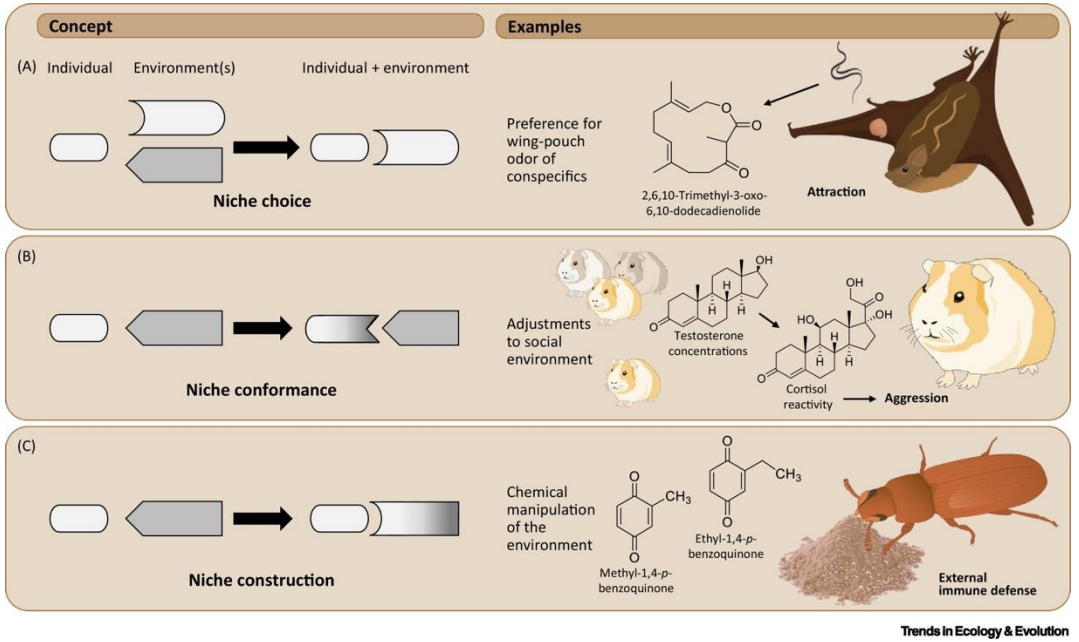
- Individuals interact with an environment to form “niches”

Main Processes (Odling-Smee, Laland, etc.):

- Niche Choice: Selecting fitting environmental conditions
- Niche Conformance: Adjusting phenotype to the environment
- Niche Construction: Modifying the environment to suit needs



Ecological Networks



From Ecology to Social Organisation:

- Individuals interact with an environment to form “niches”

Main Processes (Odling-Smee, Laland, etc.):

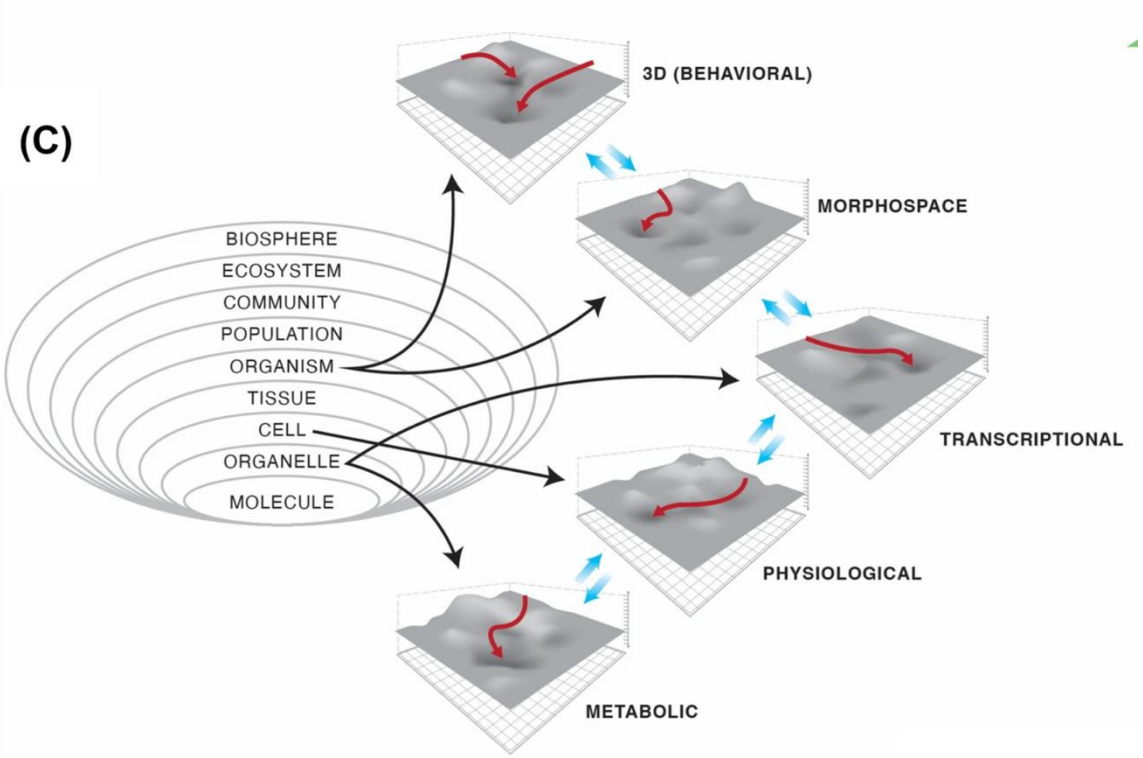
- Niche Choice: Selecting fitting environmental conditions
- Niche Conformance: Adjusting phenotype to the environment
- Niche Construction: Modifying the environment to suit needs

Application to AI:

- Agents dynamically adapt or reshape their network connections
- Mediated by targeted communication channels

Manipulating Inter-Agent Communication

Watson & Levin's Insight: Inter-cellular signalling can be "re-written" to control collective outcomes



Bioelectric networks: the cognitive glue enabling evolutionary scaling from physiology to mind

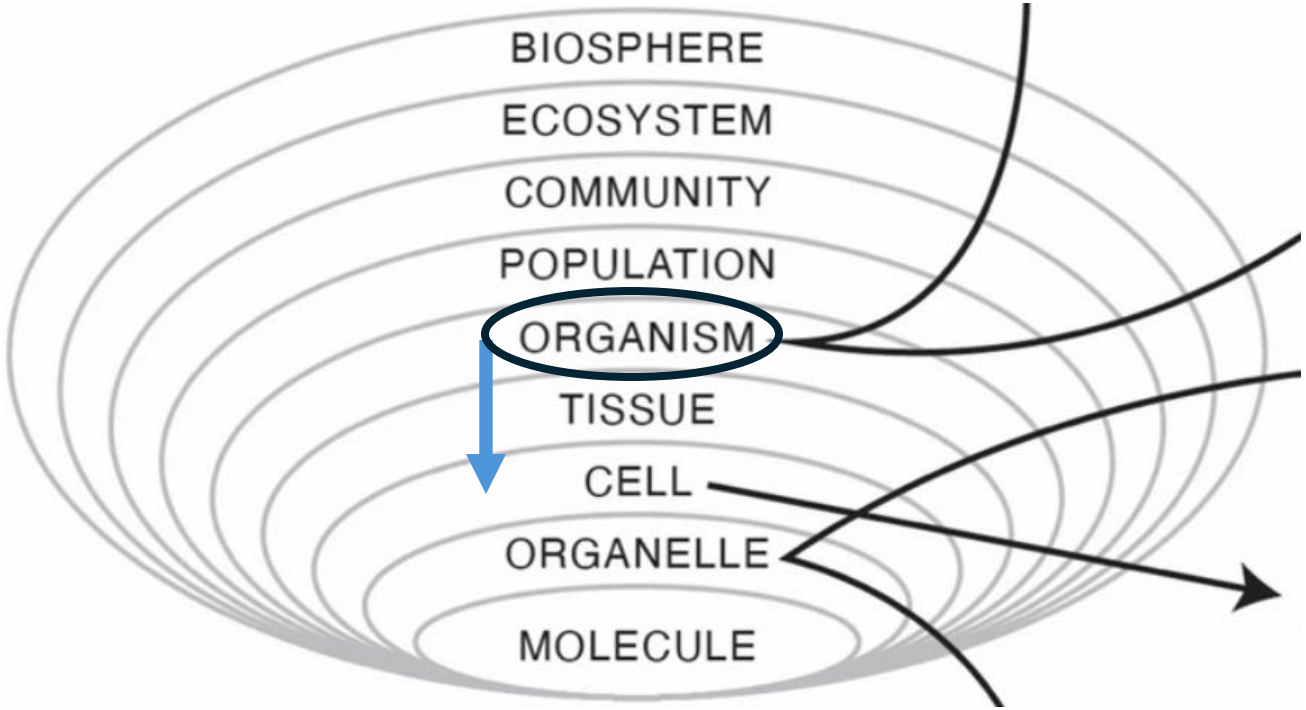
Review | Open access | Published: 19 May 2023

Volume 26, pages 1865–1891, (2023) Cite this article

Manipulating Inter-Agent Communication

Watson & Levin's Insight:

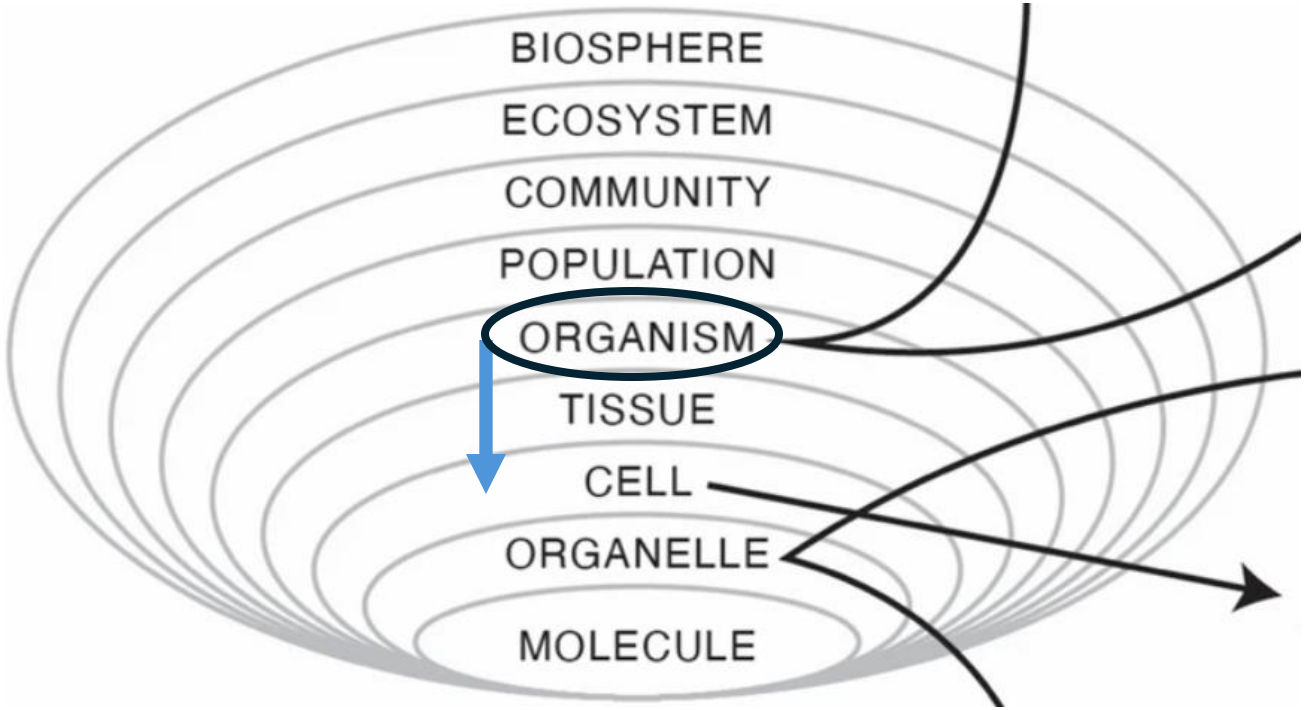
Inter-cellular signalling can be "re-written" to control collective outcomes



Manipulating Inter-Agent Communication

Watson & Levin's Insight: Inter-cellular signalling can be "re-written" to control collective outcomes

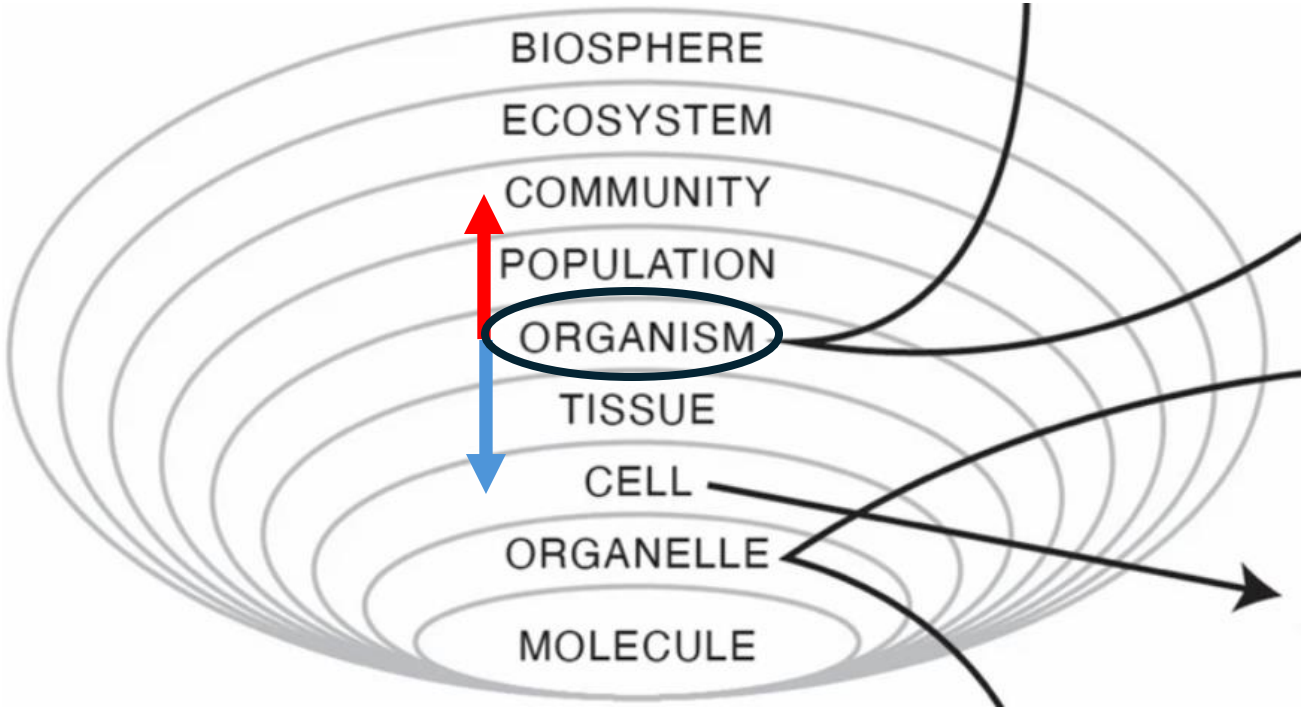
"This framework [of collective cellular intelligence] makes a strong prediction: if intercellular signalling (not genes) is the cognitive medium of a morphogenetic individual, **it should be possible to exploit the tools of behavioural and neuro-science and learn to read, interpret and re-write its information content** in a way that allows predictive control over its behaviour (in this case, growth and form) without genetic changes."



Manipulating Inter-Agent Communication

Watson & Levin's Insight:	Inter-cellular signalling can be "re-written" to control collective outcomes
Parallel in Human Social Groups:	Individuals can reshape interpersonal connections to guide group behaviour
Goal-Directed Psychology:	Agents manipulate relationships for higher-level community or population goals

"A counter question is: How can single agents, such as human leaders, have predictive control over a social group? Just as a scientist external to a cell collective can manipulate inter-cellular signalling to control the outcomes of the cell collective, a leader internal to a human collective can manipulate inter-personal behaviours to control the outcomes of the human collective."



Enter Theory of Mind (ToM)

Definition:

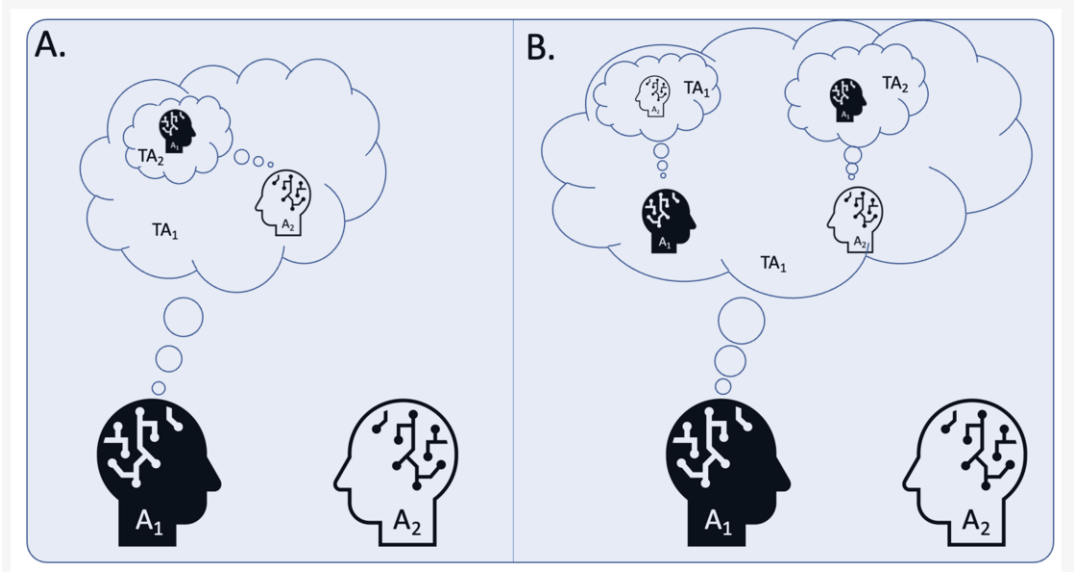
- Ability to represent and reason about others' mental states

Critical Role in Social Organisation:

- Understanding others' goals, beliefs, intentions

Darwin's "Strange Inversion" for social networks:

- Darwin: Complexity need not be consciously "created"
- We can unconsciously create complex social networks
- (or with some partial awareness)



What Can Game Theory Tell Us about an AI 'Theory of Mind'?

by Michael S. Harré

Complex Systems Research Group, Faculty of Engineering, The University of Sydney, Sydney 2006, Australia

Enter Theory of Mind (ToM)

Definition:

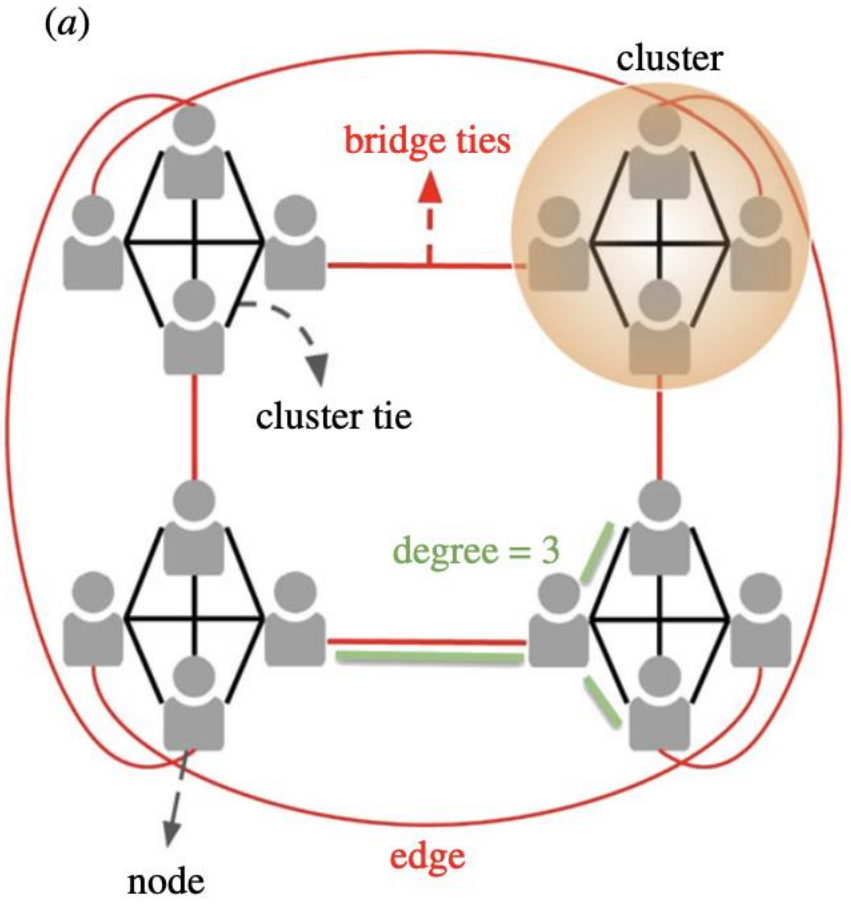
- Ability to represent and reason about others' mental states

Critical Role in Social Organisation:

- Understanding others' goals, beliefs, intentions

Darwin's "Strange Inversion" for social networks:

- Darwin: Complexity need not be consciously "created"
- We can unconsciously create complex social networks
- (or with some partial awareness)



Collective minds: social network topology shapes collective cognition

Ida Momennejad

Microsoft Research NYC, New York, NY, USA

IM, 0000-0003-0830-3973

ToM and Language

Representational View:

- Specific grammatical structures (complements) help express complex mental states

Developmental Evidence:

- Children's language acquisition strongly predicts false-belief (ToM) tasks
- Training in complement syntax enhances ToM

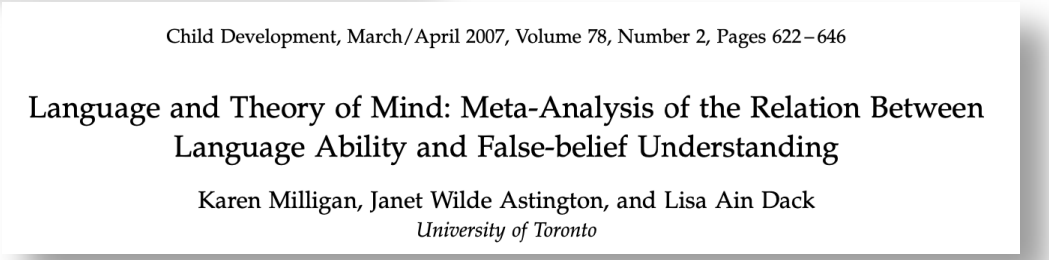
Insight for AI:

- Language-as-cognitive-tool critical for modelling others' internal states

First Language, 22 (2002), 197–213. Printed in England

Early language development and the emergence of a theory of mind*

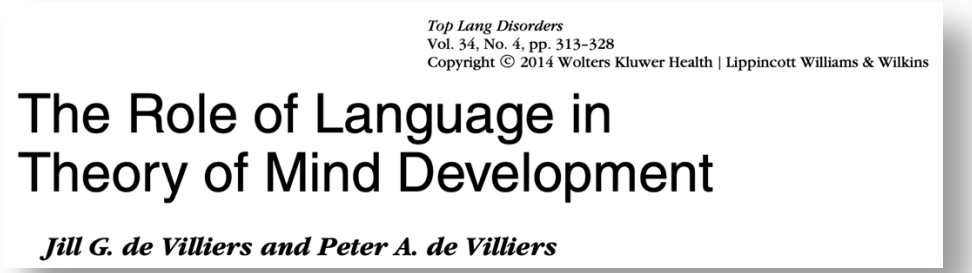
M. JEFFREY FARRAR
LISA MAAG
University of Florida



Child Development, March/April 2007, Volume 78, Number 2, Pages 622–646

Language and Theory of Mind: Meta-Analysis of the Relation Between Language Ability and False-belief Understanding

Karen Milligan, Janet Wilde Astington, and Lisa Ain Dack
University of Toronto



Top Lang Disorders
Vol. 34, No. 4, pp. 313–328

The Role of Language in Theory of Mind Development

Jill G. de Villiers and Peter A. de Villiers

Biological Theory (2023) 18:234–252
https://doi.org/10.1007/s13752-020-00372-5

THEMATIC ISSUE ARTICLE: ARCHAEOLOGY AND COGNITIVE EVOLUTION

Causal Cognition and Theory of Mind in Evolutionary Cognitive Archaeology

Marlize Lombard[1,3] · Peter Gärdenfors[1,2,3]

ToM, Language and Causal Cognition

Lombard & Gärdenfors Hypotheses:

- ToM as a crucial component of causal cognition
- More sophisticated causal cognition requires richer ToM
- Evolution of causal cognition relies on mental representations of hidden variables

Focus:

- Causal modelling of behaviour in social contexts

First Language, 22 (2002), 197–213. Printed in England

Early language development and the emergence of a theory of mind*

M. JEFFREY FARRAR
LISA MAAG
University of Florida

Child Development, March/April 2007, Volume 78, Number 2, Pages 622–646

Language and Theory of Mind: Meta-Analysis of the Relation Between Language Ability and False-belief Understanding

Karen Milligan, Janet Wilde Astington, and Lisa Ain Dack
University of Toronto

Top Lang Disorders
Vol. 34, No. 4, pp. 313–328

The Role of Language in Theory of Mind Development

Jill G. de Villiers and Peter A. de Villiers

Biological Theory (2023) 18:234–252
https://doi.org/10.1007/s13752-020-00372-5

THEMATIC ISSUE ARTICLE: ARCHAEOLOGY AND COGNITIVE EVOLUTION

Check for updates

Causal Cognition and Theory of Mind in Evolutionary Cognitive Archaeology

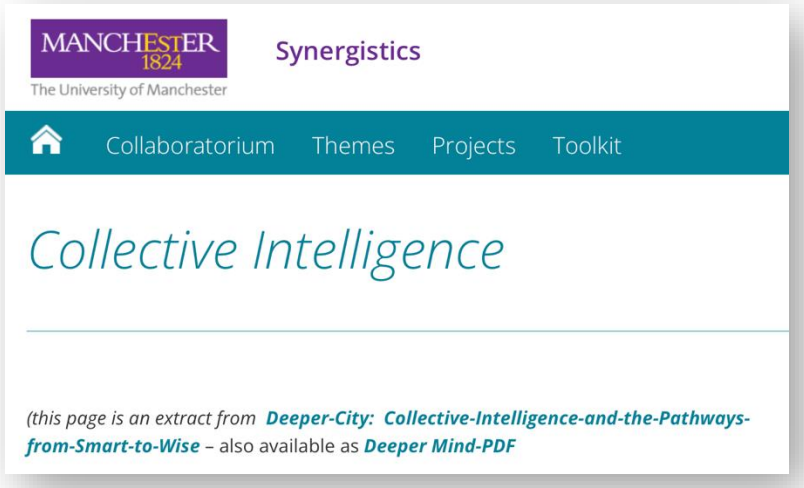
Marlize Lombard[1,3] · Peter Gärdenfors[1,2,3]

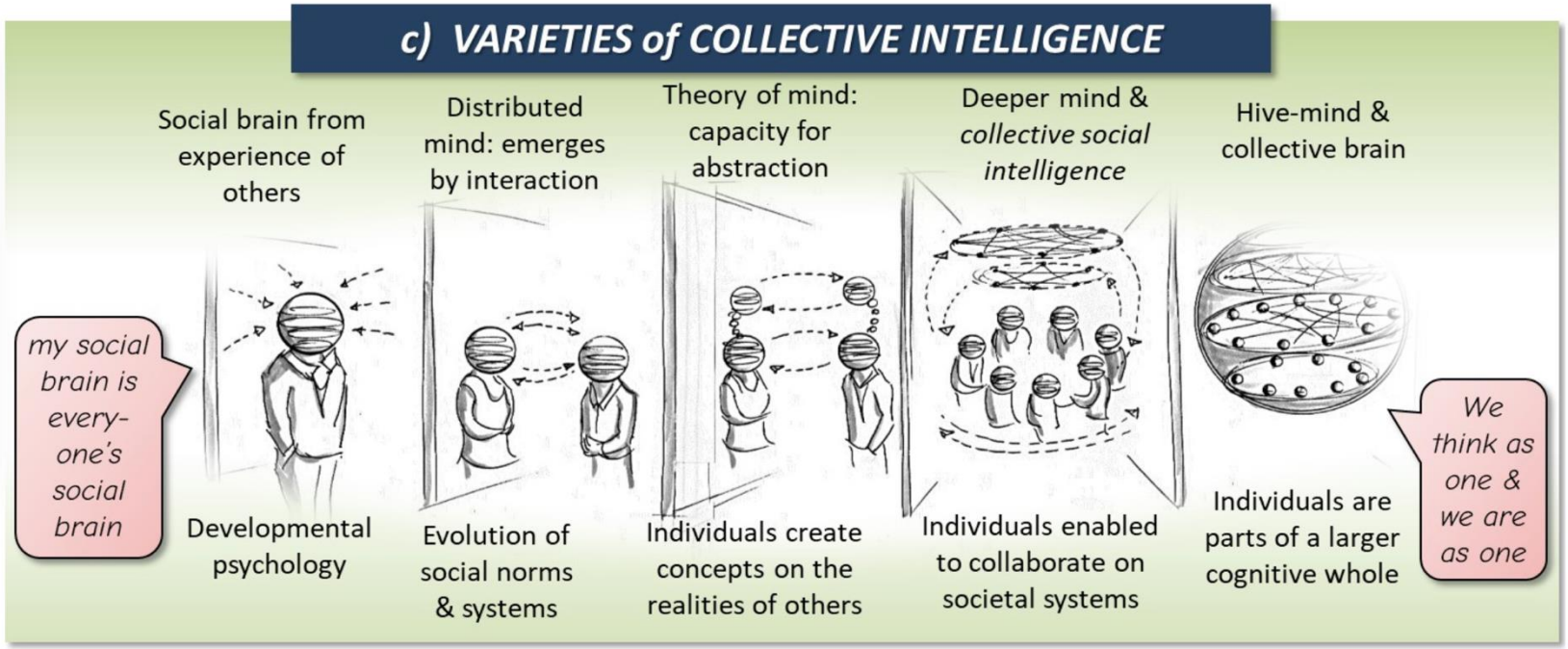
ToM & Collective Intelligence

- **Empirical Findings:**
  - Higher ToM competence → Improved group performance
- **Mechanism:**
  - Ability to read, interpret, and re-write social links
- **Social Cognitive Toolbox:**
  - Language, shared causal understanding, ToM

ToM & Collective Intelligence

- **Empirical Findings:**
  - Higher ToM competence → Improved group performance
- **Mechanism:**
  - Ability to read, interpret, and re-write social links
- **Social Cognitive Toolbox:**
  - Language, shared causal understanding, ToM

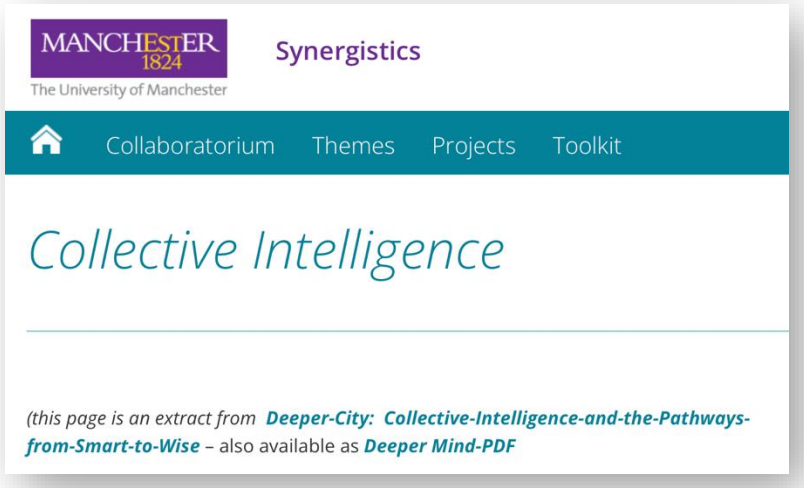


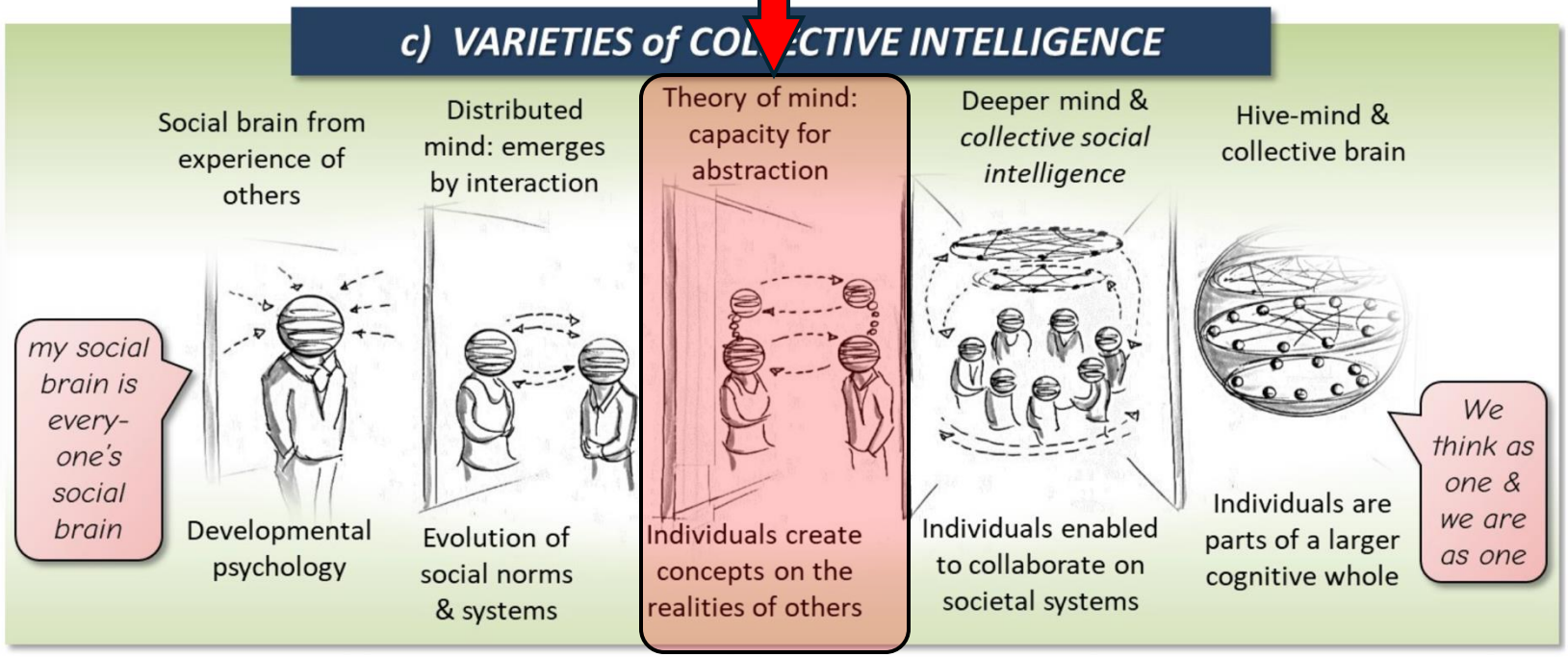


https://sites.manchester.ac.uk/synergistics/home-collective-intelligence/

ToM & Collective Intelligence

- **Empirical Findings:**
  - Higher ToM competence → Improved group performance
- **Mechanism:**
  - Ability to read, interpret, and re-write social links
- **Social Cognitive Toolbox:**
  - Language, shared causal understanding, ToM

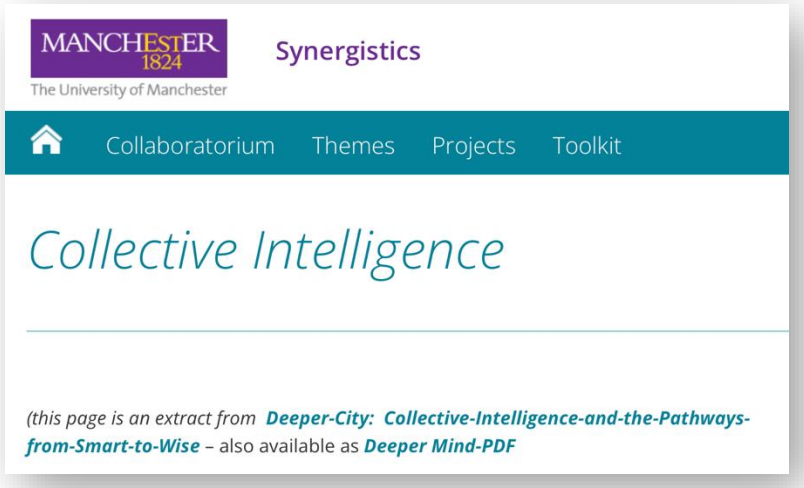


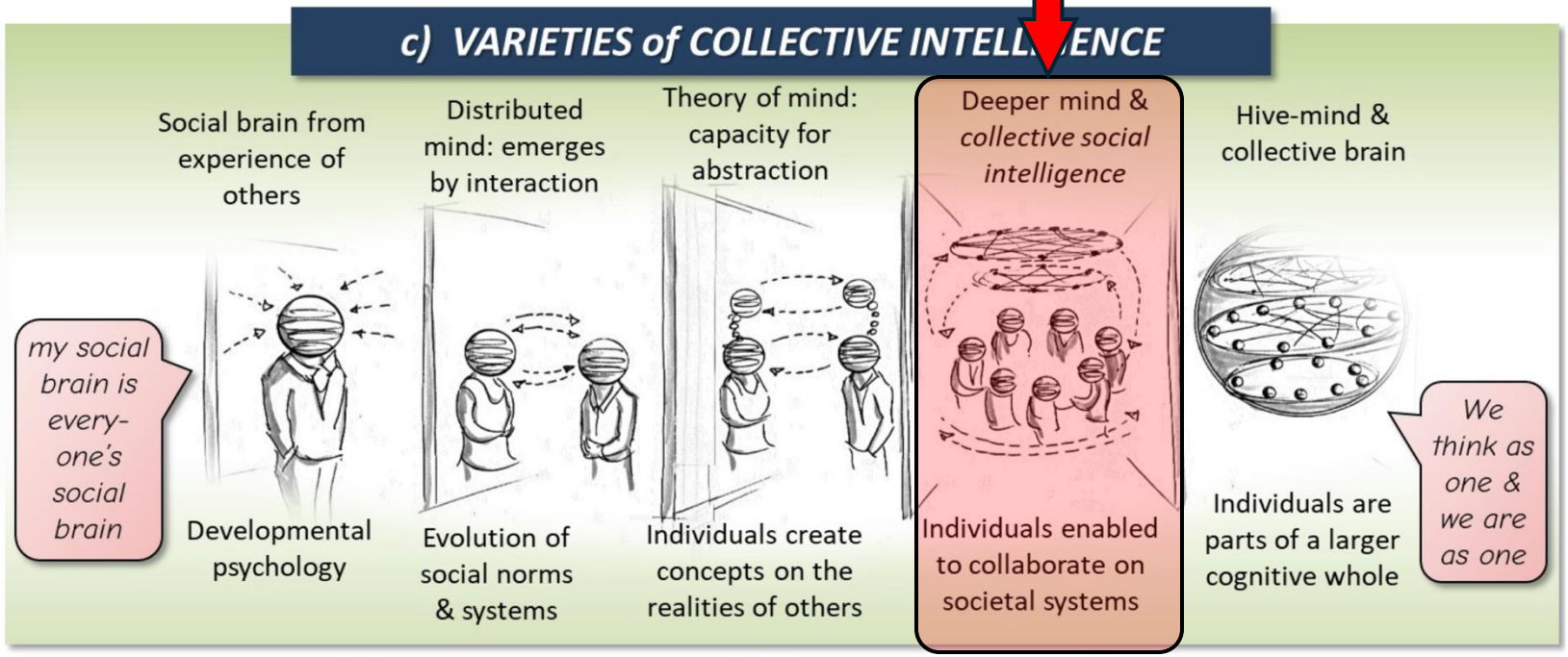


https://sites.manchester.ac.uk/synergistics/home-collective-intelligence/

ToM & Collective Intelligence

- **Empirical Findings:**
  - Higher ToM competence → Improved group performance
- **Mechanism:**
  - Ability to read, interpret, and re-write social links
- **Social Cognitive Toolbox:**
  - Language, shared causal understanding, ToM





https://sites.manchester.ac.uk/synergistics/home-collective-intelligence/

Gaps in Current AI-ToM

Inverse Reinforcement Learning (IRL):

Infers an agent's reward function from behaviour

Key Limitation: Ignores the broader social network

Large Language Models (LLMs):

Unpredictable failure modes

Mimics social cognition

No agency (planning, decision-making, goal directed behaviour),

Little feedback with a "working" environment

Limited Socially Embodied AI:

LLMs or IRL agents not situated in dynamic communication networks

Missing capacity for network manipulation based on shared causal models

Comparison to Early Humans:

Humans could reorganize social ties intentionally to achieve group goals

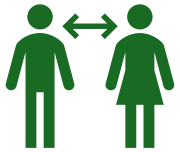
Tasks for AI-ToM

Primitive human-like skill acquisition

Lombard-Gärdenfors (2023) cognitive levels

Scaffold the learning processes for AI-ToM

How does abstraction of social interaction work?



Repeated Stochastic Game Theory:

Cheap talk relationship building in general games

Role of trust building and internal model inference?

(Crandall et al, 2018)

Complex games, more natural-like:

Cicero (AI-Meta) playing the game *Diplomacy*

Reduce Diplomacy to its core elements

When/how does deception emerge?

Collective Intelligence Learning via Info.Theory:

Collective Intelligence

Computation using Info Theory

Learning via local optimization

When Liquid vs Solid?

Review

Review

Review

Review

- Grade 1: Individual causal cognition
- Grade 2: Cued dyadic cognition
- Grade 3: Conspecific Theory of Mind
- Grade 4: Detached dyadic-causal cognition
- Grade 5: Non-conspecific Theory of Mind
- Grade 6: Inanimate causal cognition
- Grade 7: Causal network cognition

Biological Theory (2023) 18:234–252
https://doi.org/10.1007/s13752-020-00372-5

THEMATIC ISSUE ARTICLE: ARCHAEOLOGY AND COGNITIVE EVOLUTION

Causal Cognition and Theory of Mind in Evolutionary Cognitive Archaeology

Marlize Lombard[1,3] · Peter Gärdenfors[1,2,3]

Review

Grade 7: Causal Network Cognition

"**Causal network thinking** is a critical development, because it allows us to **gain new knowledge or insight from what we already know** through either individual discovery or socially transmitted knowledge **about hidden variables**."

Biological Theory (2023) 18:234–252
https://doi.org/10.1007/s13752-020-00372-5

THEMATIC ISSUE ARTICLE: ARCHAEOLOGY AND COGNITIVE EVOLUTION

Causal Cognition and Theory of Mind in Evolutionary Cognitive Archaeology

Marlize Lombard[1,3] · Peter Gärdenfors[1,2,3]

Future Outlook

Summary:

- Biological and ecological analogies illustrate “self-guided” multi-agent intelligence
- Human social cognition is a potent direction to explore in AI design

Caution:

- Avoid oversimplifying or misusing rich psychological constructs (Shevlin & Halina)

Path Forward:

- Develop AI that dynamically reconfigures social structures - harnesses fluid collective intelligence
- Integrate causal social cognition, language, and ToM for genuine self-guided organization and social learning